United Nations Industrial Development Organization

Distr.
LIMITED

ID/WG.382/2/Add.6
20 September 1982

ORIGINAL: ENGLISH

High-level Meeting on the Establishment of
the International Centre for Genetic Engineering
and Biotechnology

Belgrade, Yugoslavia, 13-17 December 1982

BIO-INFORMATICS*

prepared by

Carl-Göran Hedén**

* The views expressed in this paper are those of the author and do not necessarily reflect the views of the secretariat of UNIDO. This document has been reproduced without formal editing.

** Professor, Karolinska Institutet, Stockholm, Sweden.

V.82-30510

Contents

	Page
A. BACKGROUND AND JUSTIFICATION	1 - 8
B. ACTIVITIES	8 - 9
C. WORK PLAN	9 - 10
D. PREREQUISITES	10 - 11
E. FINANCIAL REQUIREMENTS	11 - 12
- Five-Year Budget	12
REFERENCES	13

A. BACKGROUND AND JUSTIFICATION

The term bio-informatics has been suggested for the area of interaction between information technology and the life sciences, including biotechnology. As the means for defining the fundamental intrinsic characteristics of microorganisms and other cells in numerical terms became available, this area has developed into an important service activity. This provides structural and functional information on macromolecules and metabolites and has initiated the development of mathematical models that illustrate the dynamic interactions within and between cells. It thus helps the microbial ecologist to understand the digestion processes which occur both inside and outside the body, and it also indicates how complex phenomena in soil and water might be guided towards human benefit and resource conservation. Bio-informatics also tells the fermentation engineer where to find the best strains and substrates for a given purpose, and indicates the "feed-back" strategies that are most likely to yield the desired industrial product. Finally, it provides the artificial intelligence which integrates the information from various instruments into identification labels for bacteria (numerical taxonomy), as well as diagnostic information for doctors, guidelines for biochemists and support to the many industrial practitioners of applied microbiology.

Bio-informatics gained a new dimension when it was understood that all biological processes depend on genetic information stored as linear codes along gigantic chain molecules (DNA). The codes are universal and made up of four basic units paired in a characteristic fashion. Some 300 to 2000 such pairs make up a functional piece of information, a gene. The number of genes in a cell is determined by its functional requirements: a virus might need some 35,000 pairs of the basic building blocks, a bacterium 35 million and a human cell perhaps 35 billion. This makes for great differences in the length of the information strands, as well as in the "packaging" and reading

the messages. In higher, specialized cells, most of the instructions are normally kept silent except, of course, for the little segment needed for a defined function. In fact, in higher animals the silent regions have increased in size so much that nature devised cutting and splicing procedures that transmit only selected instructions for translation into proteins, such as the molecules which accelerate the specific reactions that characterize a particular cell.

Such biocatalysts, or enzymes, function very much like the numerically controlled machining equipment in a group of factories that supply each other with tools, energy sources and prefabricated parts. In the same way as the machines perform millions of operations before they have to be replaced, each enzyme molecule can process large quantities of material before it has to be replaced. Provided with the proper raw materials (sugar, fats, etc.) and energy sources (light or energy stored in chemical bonds) the cell, however, shows a flexibility that goes far beyond anything yet seen in industry. Industry cannot, for instance, rapidly reduce the number of machines to adjust the production capacity to market needs or to resource availability. A microbial cell can do this by making use of its storehouse of information molecules, and it also uses this information to copy itself rapidly until the environment has been exhausted or until competing organisms get the upper hand. The most successful information carried by one organism, or by a combination of organisms, is thus continually selected and multiplied.

Laboratory procedures devised to increase the genetic diversity in microbial populations and to improve the efficiency of selection for useful strains have long been used by applied microbiologists. They often involve the production of "mutants" through the destruction, by irradiation or chemicals, of the control mechanisms which the cell has developed over thousands of years to guard against wastage of energy or materials. A microorganism is thus produced which can serve man well, provided that it is supplied with appropriate nutrients and is protected from competion with its relatives by means of a test tube or a fermenter.

New knowledge about the behaviour of microorganisms and other cells under stress has influenced fermentation technology, as have also the results obtained by microbial ecologists studying the stability and versatility of microbial ecosystems. Defined mixed cultur for converting natural products into useful chemicals are on the horizon, and new ways to exploit microbial enzymes are continuously being reported.

Those facts underline the need for increased efforts in the metabolic mapping of known microorganisms and for goal-oriented selection and characterization of strains isolated from natural environments. The turnover of organic molecules is particularly high in the tropics, and this ecosystem may well represent one of humanity's largest untapped natural resources when it comes to biomass utilization and biofuel management.

Even though the advances in molecular biology have been impressive in recent years, much remains to be learnt about the production plans the cell follows when it assembles large protein molecules from a small number of building blocks: amino acids. Nature also provides thought-provoking models for planned obsolescence, for storing energy and for the recycling of old building blocks into new configurations.

The feedback principles, energy-saving practices and recycling methods used in modern industry show great similarities with the approaches that nature has developed by trial-and-error methods over millions of years. However, our industrial methods are crude if compared with those that are used by the living cell. We are only now starting to understand that nature's method of letting a sequence of reactions dovetail into each other over surfaces, where the micro-environment for each step is neatly adjusted, might eventually help the chemical industry to save a lot of expensive centrifuges, filters. and flocculation tanks. Nature's preference for continuous rather than batch processing, and its capacity to do without corrosive chemicals

and high temperatures and pressures, further indicate how fuel, burners and heat exchangers might perhaps be saved, and how expensive tanks, pumps and valves might be replaced with simple equipment made from glass or plastic. Furthermore, the reduced vulnerability derived from operating on a smaller scale and with a wider spectrum of raw materials might well offset any sacrifices made in processing speed.

The enzymes which microbial cells use to break down large molecules were early used in industrial processes, but for synthesizing purposes the metabolic processes of whole cells had to be used, first in regular fermention processes and more recently in immobilized systems where the growth of the cells is restricted but their metabolism is largely intact. However, better understanding of the ways in which cells precisely adjust the energy inputs into synthesizing processes is now opening the door to new approaches based on defined cell components. These might involve either membrane-bound enzymes or the special molecules (co-enzymes) that transfer electrons between cell locations where energy is either used or stored until needed. Some years ago it was found that electricity could charge some of the transfer molecules with high efficiency and also that certain important energy-rich compounds (ATP) could be made synthetically. Those facts, as well as the observation that some biosynthetic reactions are favoured when enzymes operate in abnormal solvent environments, indicate many new opportunities for industrial processes. The fact that substances insoluble in water can be attacked by microbial cells, even when the microbes cannot multiply, for instance opens up a whole new field for biosynthesis. As an example, cholesterol can be oxidized by certain microbial cells even when they are suspended in the organic solvent carbon tetrachloride. The fact that cheap substrates, like sugar, can be used to regenerate the co-enzymes that drive the basic life processes also makes it possible to modify expensive substrates by metabolic "hitch-hiking" on properly selected microbial cells.

The knowledge gained about the function of biocatalysts may eventually lead to the synthesis of molecules ("synzymes") with a higher stability than that which is desirable for a living cell. Recent

advances in the immobilization of enzymes on electrode surfaces also make it reasonable to assume that ways will be found to open chemical bonds in such a fashion that electrons can be drawn off efficiently as an electric current. Such "biochemical fuel cells" might offer new approaches to the decentralized production of electricity from such easily available materials as alcohols, methane and ammonia. They would also offer novel approaches to the development of new types of environmental sensors.

It is obvious that opportunities such as those just outlined will be under constant review by the International Centre for Genetic Engineering and Biotechnology's (ICGEB's) Council of Scientific Advisers and by the heads of the departments that need to make use of the specialized facilities and the library managed by the Bio-informatics Department. This will act not only as a culture collection, but also as a storehouse of metabolic information about strains and as a sequence library and switchboard to data banks holding detailed information about amino acids and nucleotides. It will also be responsible for ICGEB's mainframe computer and serve as a major programming and computer science resource for the Centre. Besides having a service function, it would also have its own research profile, which might well be geared to the development of programmes specially adapted to the three-dimensional presentation of the various molecules that interest genetic engineers and molecular biologists. The vibrating mirror concept recently invented by Dr. Laurence Shea (1) for instance offers new possibilities to vizualize the fine structure of the active sites in complex organic molecules. It is an example of equipment that is most efficiently used if it forms part of a major research establishment.

Highly specific sensors using enzymes to detect and measure small organic molecules already have medical and industrial applications. They exemplify the type of instrumentation which is now continuously broadening the data base of bio-informatics. Cumbersome physical techniques for the study of large molecules have also been supplemented

with immunological methods based on the use of pure antibodies (the body's highly specific defense molecules) made by fusing tumour cells that can grow very fast with other cells having the capacity to produce antibodies that bind to one trigger antigen only. Such "monoclonal antibodies" can be used not only to detect and measure very large molecules; they also open new industrial approaches to "drug targeting" and for immunotherapy. The latter can be considered in cases (e.g. cancer, rabies, some parasitic diseases) where the benefits might outweigh the small, but conceivable, risk that the preparation might carry a cancer virus. For the biochemist the greatest attraction of such antibodies is, however, that they can be used to "fish out" specific molecules from crude extracts of cells that manufacture desired products. Pure material can thus be obtained for analysis and for use of a model in the synthesis of the corresponding string of genetic information. This can be done by various techniques, some of which depend on the use of sophisticated laboratory robots.

13. In order to make full use of the potential of monoclonal antibodies it is important for the ICGEB to build up and maintain a collection of appropriate cell lines. The low-temperature refrigeration equipment required for this operation can also be used for the strain and vector collection which should be started as soon as possible, since it could provide a valuable service to the laboratories associated with the ICGEB.

14. Sequence libraries are gradually developing. They provide predictions about the positions where specific enzymes are likely to cut the DNA chain, indicate how molecular "bait" should look in order to "fish out" desired information strands from disrupted cells, and finally, guide the construction of genes that will be effective in a particular microorganism.

15. In view of the significance of such libraries (supplemented with the long-existing libraries of amino acid sequences), and of collections containing cultures of cells that produce monoclonal antibodies or vectors (a ring of DNA called a plasmid, or a virus) selected for their capacity to "package" and store large DNA fragments,

more international co-ordination is needed. The knowledge and materials held in such libraries and collections represent a critically important tool for gaining understanding about tumour development, metabolic diseases and autoimmune reactions, and they can also stimulate industrial microbiology.

Suitably modified synthetic or isolated genes can be attached to an appropriate vector, and a particle is thus formed which has the capacity not only to penetrate a microbial cell but also to force it to read the information provided and to translate it into large quantities of a desired product. The amount produced depends on the efficiency of the reading system and on the number of copies of genetic information available to the cell, now often increased by a technique referred to as gene amplification. The fermentation engineer can thus be provided with highly productive strains, with strains that make molecules characteristic of higher organisms and with hybrid strains that combine desirable properties from very different micro-organisms. Obviously, the storage and distribution of such strains would represent the management of packages of bio-informatics of great relevance to the development process in poor countries.

As indicated before, a Bio-informatics Department would have an important internal service function for ICGEB, supplying materials, equipment and expert knowledge in the computer field. However, it would also have an important "outreach" function to associated laboratories and to national centres in developing countries.

Fortunately, communication technology has now reached a point where an effective international pooling of resources is possible. Via the commercial network TYMNET a scientist in Europe can for instance easily connect to the American SUMEX system (3), which contains a data bank that holds a large number of sequences from various organisms and viruses, as well as for immunoglobulins, t-RNA's, satellites, replication origins, globins, transposable elements, caps, etc. He can also find the first enzyme discovered for each nucleotide sequence cleaved, and enter his own data in order to generate restriction maps.

However, many other data processing networks, like PROPHET (4), may need to be culled, so it is obvious that the Bio-informatics Department must be able to provide terminal guidance. This type of training will be important for scientists coming from developing countries. They will learn to appreciate the importance of data banks and will be able to benefit from electronic mail servicing as an umbilical cord when they have returned home. Finally, they may then continue to participate in international research co-operation by means of computer conferencing (5).

B. ACTIVITIES

(a) Computer simulation of biological activities.

The technique for determining the exact sequence of building blocks that makes up a gene has developed very rapidly in recent years. Thanks to computer science and advanced instrumentation, it is now possible to "sequence" DNA at a rate of at least 200 pairs a day. Since this corresponds to a polypeptide made up of 67 amino acids it follows that the gene which codes for the average polypeptide (molecular weight around 35,000) can be mapped in a few weeks time. By back-reading such information, the computer can also predict the structure of the messenger molecule that transmits the instruction from the gene to the protein building site in the cell. It can also predict the amino acid sequence of the polypeptide that will eventually emerge. However, the way in which the long nucleotide string folds is critically important for its function, so computer-aided predictions of the secondary structure are very important. For RNA two approaches can then be used (2). One utilizes a thermodynamic energy minimization method that takes into account the likelihood that short-range folding tends to be favoured over long-range interactions. The other method utilizes interactive computer graphics modelling that enables the user to consider thermodynamic criteria as well as structural data obtained by nuclease susceptibility, chemical reactivity and phylogenetic studies. These two approaches are now being combined so that the NIH user can start with a candidate hairpin structure, generated by the thermodynamic approach, and at any point in his interactive refinement

of the structure he will be able to invoke the thermodynamic algorithm on any piece of the molecule that he selects. The power of such an approach is obvious, but so is the need for expert management.

(b) Development of software requirement for ICGEB.

(c) Documentation and dissemination of information.

(d) Organization of culture and vectors collection.

(e) Support services.

C. WORK PLAN

Year 1:

- Establishment of research priorities by the Council of Scientific Advisers;
- Ordering of mainframe computer and terminal equipment for ICGEB's departments and associated institutions;
- Preparation of networking with relevant data banks;
- Development of microfiche laboratory and programme libraries with special emphasis on computer graphics applied to molecular biology and to optimization in bio-engineering;
- Organization of library and documentation services.

Year 2:

- Establishment of collection of strains and vectors;
- First central course in bio-informatics;
- Provision of support services to other programmes.

Year 3:

- First expert meeting and seminar;
- Second central course;
- Selection of five new trainees for associated laboratories;
- Provision of support services to other programmes.

Year 4:

- Provision of support services to other programmes.

Year 5:

- Third expert meeting and seminar;
- Provision of support services to other programmes.

D. PREREQUISITES

Two keys are necessary to fully open the treasure chest of genetic diversity represented by the microbial kingdom, which actually makes up one quarter of the total weight of all living matter (both plants and animals) in the world. One is the hardware of bio-informatics: information about how the analytical techniques, biosynthetic methods and genetic manipulations referred to above are performed. It also includes information about the availability of the microorganisms, vectors, chemicals and equipment needed. Examples are the computerized abstracting services and patent files, the World Data Bank on Micro-organisms (Brisbane) and all the various lists of commercially available instruments, chemicals and enzymes that the microbial geneticist uses as tools to copy, cut and splice genetic information. Basically this is the ICGEB library and internal data base.

The other key is the software of bio-information: the quantitative metabolic information and analytical data that characterize various cells and organic molecules, and the mathematical models which illustrate their interactions. Examples are the computer programmes used to analyse the X-ray diffraction patterns which reveal the

structure of large molecules; the mass spectrometer signals that sort out the composition of complex mixtures of chemicals; and the dynamics which explain ecological interactions and the behaviour of fermentations. A gigantic effort is needed to sift and supplement such information so that the full impact of biomolecular engineering can be realized.

E. FINANCIAL REQUIREMENTS

The bio-informatic programme will dispose of personnel at various levels of expertize:

- one senior scientist;
- two junior scientists; and
- two technicians.

In addition, two computer experts (listed as supporting personnel in the budget document) will work with the above scientist.

The budget for the five-year initial period of the ICGEB is presented in the following table.

Five-Year Budget

STAFF

(First year 40 per cent, second year 60 per cent of full operation.)

		(US$ thousand)
Senior Scientist	4 man years	300
Junior Scientist	8 man years	360
Technicians	8 man years	136
Subtotal		796
Management of the Centre and Supporting Personnel		205
Total Staff		1,001
OPERATIONAL ACTIVITIES		
Expert Group Meetings	2	50
Information material		600
Associateship		75
Miscellaneous (travel, telephone, postage, etc.)		30
Total Operational Activities		755
TOTAL WORK PROGRAMME		1,756

REFERENCES

1. Peterson, I. "Mirroring a three-dimensional world": Science News Weekly. For further information contact Dr. Laurence Shea, Computer Systems Division, Bolt Beranetz and Neuman Inc. (50 Moulton Street, Cambridge, Mass. 02238) or Genisco Technology Corp., California, 1982.

2. Auron, P.E., Rindone, W.P., Vary, C.P.H., Celentano, J.J. and Vournakis, J.N. "Computer-aided prediction of RNA secondary structures", Nucleic Acids Research, 10, No.1, page 403, 1982.

3. Friedland, P. Brutlag, D. and Kedes, L. "An introduction for guests of MOLGEN on SUMEX-AIM. Dept. of Computer Science, Standford University and Depts. of Biochem. and Med., Stanford University School of Medicine, July 1, 1980.

4. Editorial: "The lab notebook is a computer". Chemical Week, Feb. 20, 1980.

5. "Bioconversion of lignocellulose for fuel, fodder and food needed for rural development in poor countries". First announcement about an international computer conference organized by the World Academy of Art and Science (WAAS) and the International Development Research Centre (IDRC) in co-operation with the UNEP/UNESCO/ICRO Microbial Panel, UNIDO and the International Federation of Institutes for Advanced Study (IFIAS). Further information from C.-G. Hedén, Karolinska Institutet, S-104 05 Stockholm, Sweden.